

UVA CS 4501 - 001 / 6501 – 007

Introduction to Machine Learning and Data Mining

Lecture 1 : Logistics & Intro

Yanjun Qi / Jane

University of Virginia
Department of
Computer Science

8/26/14

1

Welcome

- CS 4501 - 001; cross-listed as 6501 - 007
- Introduction to Machine Learning and Data Mining

- TuTh 3:30pm-4:45pm, **Thornton Hall E316**
- Course Website
 - <http://www.cs.virginia.edu/yanjun/teach/2014f/>
 - Uva Collab course page for homework submissions

8/26/14

2

Today

- Course Logistics**
- My background
- Basics of machine learning
- Application and History of MLDM

Course Staff

- Instructor: Prof. Yanjun Qi
 - Qi: /ch ee/
 - You can call me professor “Jane”
- TA: Nicholas Janus, (ncj2ey@virginia.edu)
- TA: Beilun Wang (bw4mw@virginia.edu)
- TA office hours: Monday 4:00-6:00pm @ Rice 504
- My office hours: Grab me right after a lecture

Course Logistics

- Course email list has been setup. You should have received emails already !
- Policy, the grade will be calculated as follows:
 - Assignments (60%, **SIX** total, each 10%)
 - mid-term (20%)
 - Final exam (20%)

Course Logistics

- Midterm: Oct 16, one hour in class
- Final exam: Dec 9, two hours (**tentative**)
- Six assignments (each 10%)
 - Due Sept 16, Sept 30, Oct 14, Nov 4, Nov 18, Dec 2
 - For homework-6
 - 4501-001 programming ;
 - 6501-007 course mini-project;
 - **three** extension days policy (check course website)

Course Logistics

- Policy,
 - Homework should be submitted electronically through [UVaCollab](#)
 - Homework should be finished individually
 - Due at the **beginning of class** on the due date

 - In order to pass the course, the average of your midterm and final must also be "pass".

Course Logistics

- Recommended books for this class is:
 - Elements of Statistical Learning, by Hastie, Tibshirani and Friedman. (Book PDF available online)
 - Pattern Recognition and Machine Learning, by Christopher Bishop.
- My slides – **if not mentioned in my slides, it is not an official topic of the course**

Course Logistics

- **Background Needed**

- Calculus and Basic linear algebra.
- Statistics is recommended.
- Students should already have good programming skills, i.e. 2150 as prerequisite.

- We will review “linear algebra” and “probability” in class

Today

- Course Logistics
- My background**
- Basics of machine learning & Application
- Application and History of MLDM

About Me

- Education:
 - PhD from School of Computer Science, Carnegie Mellon University (@ Pittsburgh, PA) in 2008
 - BS in Department of Computer Science, Tsinghua Univ. (@ Beijing, China)
 - My accent **PATTERN** : /l/, /n/, /ou/, /m/
- Research interests:
 - **Machine Learning, Data Mining, Biomedical Informatics**

8/26/14

11

About Me

- Five Years' of Industry Research Lab in the past :
 - 2008 summer – 2013 summer, **Research Scientist in IT** industry (Machine Learning Department, NEC Labs America @ Princeton, NJ)
 - 2013 Fall – Present, **Assistant Professor**, Computer Science, UVA



Industry + Academia

8/26/14

12

Today

- Course Logistics
- My background
- Basics of MLDM**
- Application and History of MLDM

OUR DATA-RICH WORLD



- **Biomedicine**
 - Patient records, brain imaging, MRI & CT scans, ...
 - Genomic sequences, bio-structure, drug effect info, ...
- **Science**
 - Historical documents, scanned books, databases from astronomy, environmental data, climate records, ...
- **Social media**
 - Social interactions data, twitter, facebook records, online reviews, ...
- **Business**
 - Stock market transactions, corporate sales, airline traffic, ...
- **Entertainment**
 - Internet images, Hollywood movies, music audio files, ...

Yanjun Qi / UVA CS 4501-01-6501-07

BIG DATA CHALLENGES

- Data capturing (sensor, smart devices, medical instruments, et al.)
- Data transmission
- Data storage
- Data management
- High performance data processing
- Data visualization
- Data security & privacy (e.g. multiple individuals)
-

← e.g. cloud computing

← e.g. HCI

this course

- Data analytics
 - How can we analyze this big data wealth ?
 - E.g. Machine learning and data mining

8/26/14 15

Yanjun Qi / UVA CS 4501-01-6501-07

Drowning in data, Starving for knowledge



8/26/14 16

BASICS OF MACHINE LEARNING

- “The goal of machine learning is to build computer systems that can **learn and adapt from their experience.**” – Tom Dietterich
- “**Experience**” in the form of available **data examples** (also called as instances, samples)
- Available examples are described with properties (**data points in feature space X**)

8/26/14

17

e.g. SUPERVISED LEARNING

- Find function to map **input** space X to **output** space Y $f : X \rightarrow Y$
- So that the **difference** between y and $f(x)$ of each example x is small.

e.g.

x	I believe that this book is not at all helpful since it does not explain thoroughly the material . it just provides the reader with tables and calculations that sometimes are not easily understood ...
----------	--

y	-1
----------	----

Output Y: {1 / Yes , -1 / No }
e.g. Is this a positive product review ?

Input X : e.g. a piece of English text

8/26/14

18

e.g. **SUPERVISED** Linear Binary Classifier Yanjun Qi / UVA CS 4501-01-6501-07

$\mathbf{x} \rightarrow \boxed{f} \rightarrow \mathbf{y}$

$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w}\mathbf{x} + b)$

● denotes +1 point
 ■ denotes -1 point
 ? denotes future points

8/26/14 Prof. Andrew Moore's slides ¹⁹

Basic Concepts Yanjun Qi / UVA CS 4501-01-6501-07

- **Training** (i.e. learning parameters \mathbf{w}, b)
 - Training set includes
 - available examples $\mathbf{x}_1, \dots, \mathbf{x}_L$
 - available corresponding labels y_1, \dots, y_L
 - Find (\mathbf{w}, b) by minimizing loss (i.e. difference between y and $f(\mathbf{x})$ on available examples in training set)

$$(\mathbf{w}, b) = \underset{\mathbf{w}, b}{\text{argmin}} \sum_{i=1}^L \ell(f(\mathbf{x}_i), y_i)$$

8/26/14 20

Basic Concepts

Yanjun Qi / UVA CS 4501-01-6501-07

- **Testing** (i.e. evaluating performance on “future” points)
 - Difference between true y_i and the predicted $f(x_i)$ on a set of testing examples (i.e. *testing set*)
 - Key: example x_i not in the training set
- **Generalisation**: learn function / hypothesis from **past data** in order to “explain”, “predict”, “model” or “control” **new data** examples

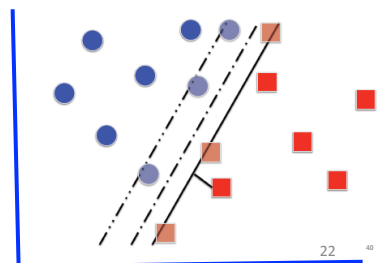
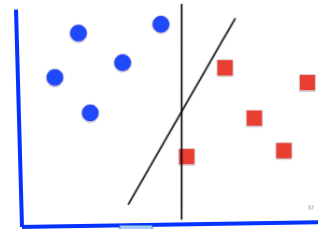
8/26/14

21

Basic Concepts

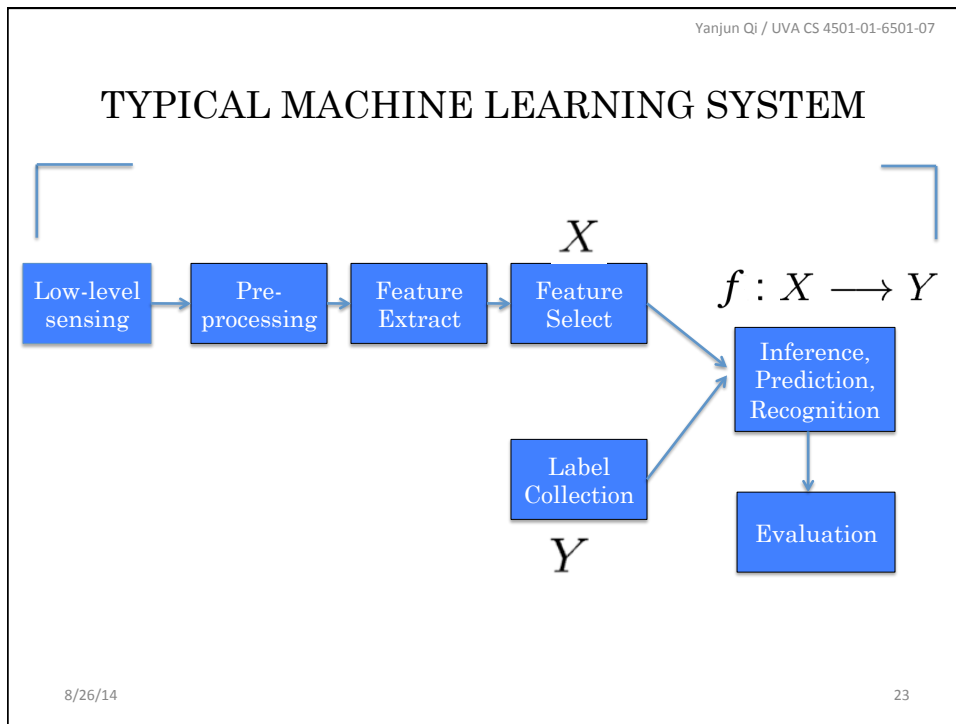
Yanjun Qi / UVA CS 4501-01-6501-07

- **Loss function**
 - e.g. hinge loss for binary classification task
 - e.g. pairwise ranking loss for ranking task (i.e. ordering examples by preference)
- **Regularization**
 - E.g. additional information added on loss function to control model



8/26/14

22



Yanjun Qi / UVA CS 4501-01-6501-07

“Big Data” Challenges for Machine Learning

LARGE-SCALE

HIGH-COMPLEXITY

- ✓ Large size of samples
- ✓ High dimensional features

Not the focus here, will be covered in advanced grad-level course next semester


8/26/14

24

Yanjun Qi / UVA CS 4501-01-6501-07

Large-Scale Machine Learning: SIZE MATTERS

LARGE-SCALE



Those are not different numbers,
those are different mindsets !!!


8/26/14

- One thousand data instances
- One million data instances
- One billion data instances
- One trillion data instances


Yanjun Qi / UVA CS 4501-01-6501-07

BIG DATA CHALLENGES FOR MACHINE LEARNING

LARGE-SCALE



Highly Complex



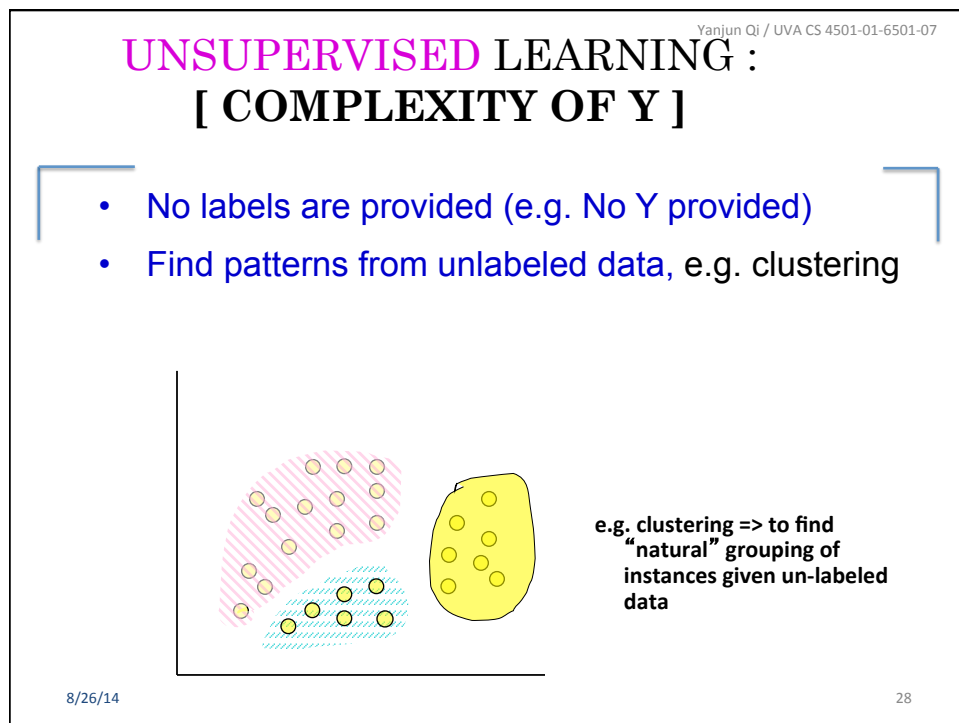
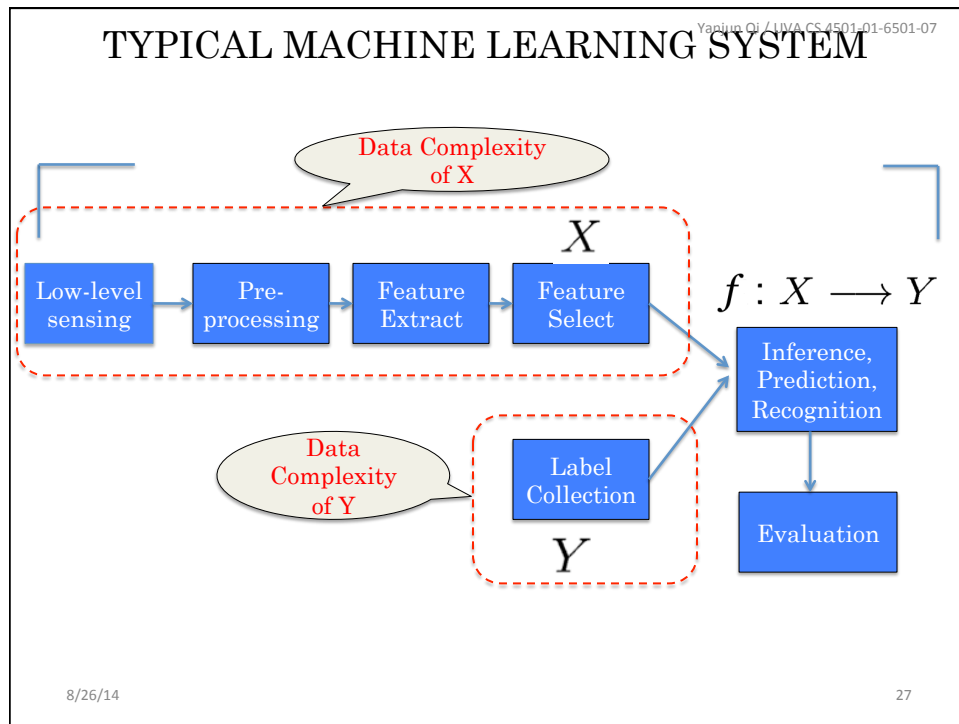
Most of
this
course

The situations / variations of both **X (feature, representation)** and **Y (labels)** are complex !

- ✓ Complexity of X
- ✓ Complexity of Y

8/26/14

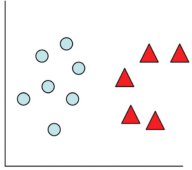
26



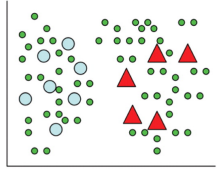
Yanjun Qi / UVA CS 4501-01-6501-07

SEMI-SUPERVISED LEARNING : [COMPLEXITY OF Y]

- Labeled data (x,y) are often **hard** to obtain
- Unlabeled data (x only) are often **easy** to obtain : **A Lot**

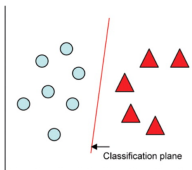


Labeled Data
(a)

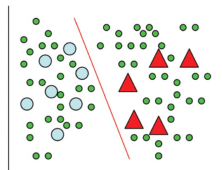


Labeled and Unlabeled Data
(b)

Combine both labeled, weakly labeled, and unlabeled examples to learn the function



Supervised Learning
(c)

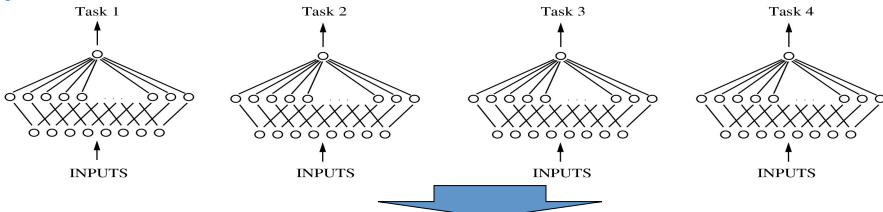


Semi-Supervised Learning
(d)

29

Yanjun Qi / UVA CS 4501-01-6501-07

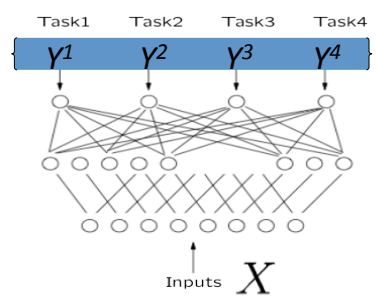
MULTITASK LEARNING: [COMPLEXITY OF Y]



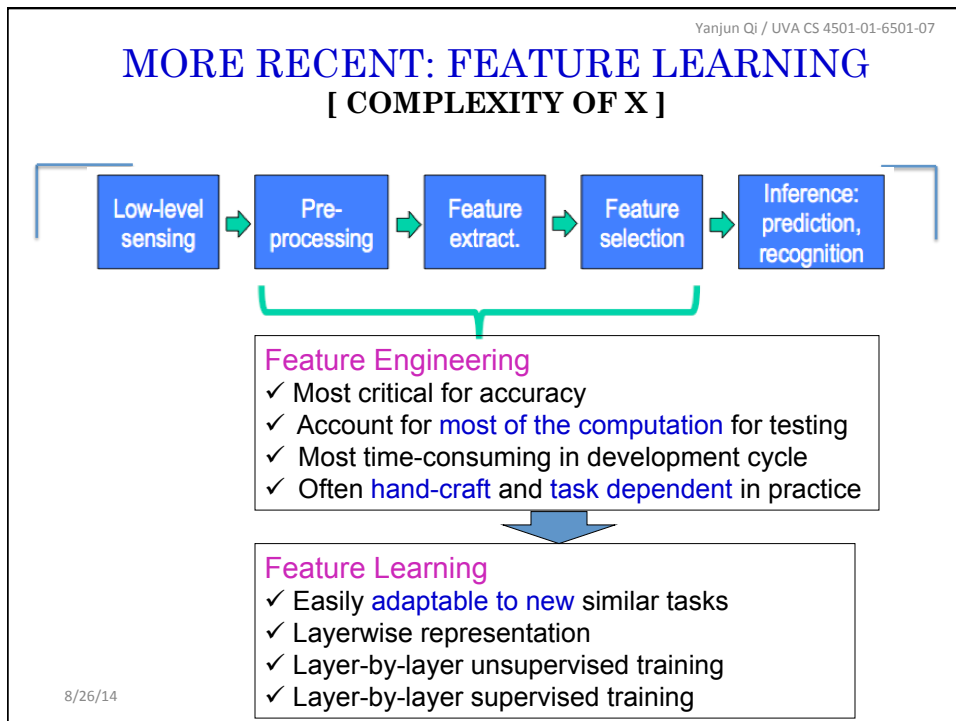
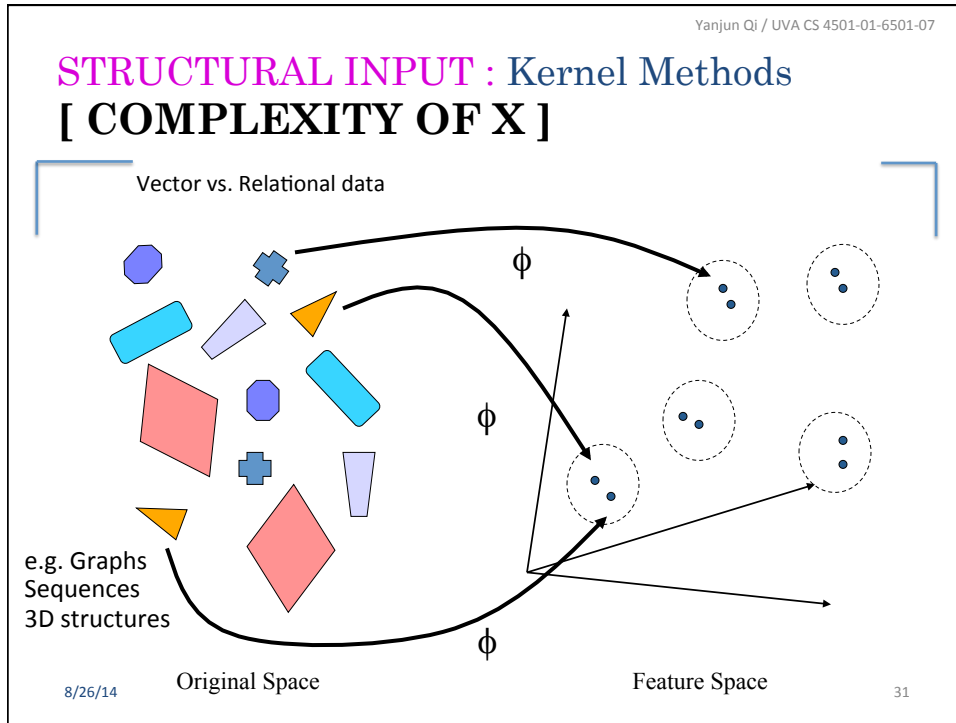
To learn a joint model

✓ Multiple closely related tasks

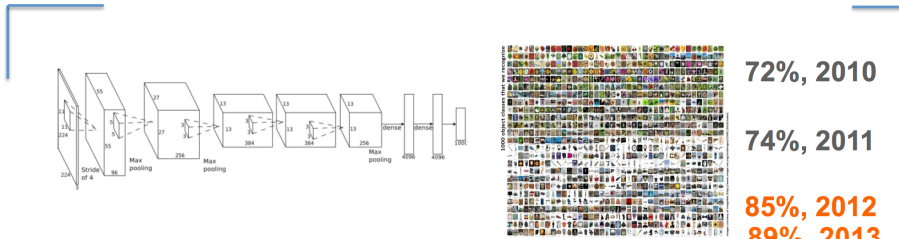
✓ Bear similar feature representations



30



DEEP LEARNING / FEATURE LEARNING : [COMPLEXITY OF X]



Deep Convolution Neural Network (CNN) just won (as Best systems) on “very large-scale” ImageNet competition 2012 and 2013

(training on 1.2 million images [X] vs. 1000 different word labels [Y])

- 2013, Google Acquired Deep Neural Networks Company headed by Utoronto “Deep Learning” Professor Hinton
- 2013, Facebook Built New Artificial Intelligence Lab headed by NYU “Deep Learning” Professor LeCun

Prof. Hinton’s slides

Today

- Course Logistics
- My background
- Basics of machine learning & Data Mining
- Application and History of MLDM

MACHINE LEARNING IN COMPUTER SCIENCE

- Machine learning is already the preferred approach for
 - Speech recognition, natural language processing
 - Computer vision
 - Medical outcome analysis
 - Robot control ...
- Why growing ?
 - Improved learning algorithms
 - Increased data capture, new sensors, networking
 - Systems/Software too complex to control manually
 -

Terminology: Some (Near-)Synonyms

- Machine learning
- Data mining
- Pattern recognition
- Computational statistics
-

Some bigger concepts that ML is part of:

- Statistics (e.g. includes hypothesis testing)
- Data analysis (e.g. includes visualization)
- Artificial intelligence (e.g. includes planning)
- Applied mathematics, computational science (e.g. includes optimization)

HISTORY OF MACHINE LEARNING

- 1950s
 - Samuel's checker player
 - Selfridge's Pandemonium
- 1960s:
 - Neural networks: Perceptron
 - Pattern recognition
 - Learning in the limit theory
 - Minsky and Papert prove limitations of Perceptron
- 1970s:
 - Symbolic concept induction
 - Winston's arch learner
 - Expert systems and the knowledge acquisition bottleneck
 - Quinlan's ID3
 - Michalski's AQ and soybean diagnosis
 - Scientific discovery with BACON
 - Mathematical discovery with AM

HISTORY OF MACHINE LEARNING (CONT.)

- 1980s:
 - Advanced decision tree and rule learning
 - Explanation-based Learning (EBL)
 - Learning and planning and problem solving
 - Utility problem
 - Analogy
 - Cognitive architectures
 - Resurgence of neural networks (connectionism, backpropagation)
 - Valiant's PAC Learning Theory
 - Focus on experimental methodology
- 1990s
 - Data mining
 - Adaptive software agents and web applications
 - Text learning
 - Reinforcement learning (RL)
 - Inductive Logic Programming (ILP)
 - Ensembles: Bagging, Boosting, and Stacking
 - Bayes Net learning

8/26/14

Prof. Mooney's slides⁴¹

HISTORY OF MACHINE LEARNING (CONT.)

- 2000s
 - Support vector machines
 - Kernel methods
 - Graphical models
 - Statistical relational learning
 - Transfer learning
 - Sequence labeling
 - Collective classification and structured outputs
 - Computer Systems Applications
 - Compilers
 - Debugging
 - Graphics
 - Security (intrusion, virus, and worm detection)
 - Email management
 - Personalized assistants that learn
 - Learning in robotics and vision

8/26/14

Prof. Mooney's slides⁴²

HISTORY OF MACHINE LEARNING (CONT.)

- 2010s
 - Speech translation, voice recognition (e.g. SIRI)
 - Google search engine uses numerous machine learning techniques (e.g. grouping news, spelling corrector, improving search ranking, image retrieval,
 - 23 and me (scan sample of person genome, predict likelihood of genetic disease, ...)
 - IBM watson QA system
 - Machine Learning as a service (e.g. google prediction API, bigml.com,
 - IBM healthcare analytics
 -

8/26/14

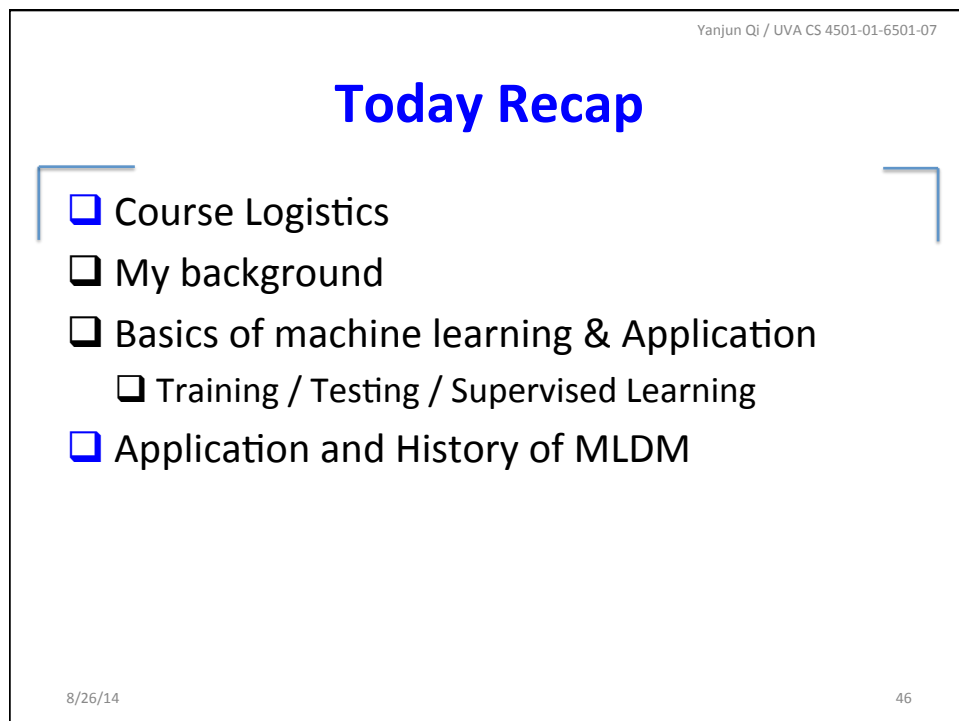
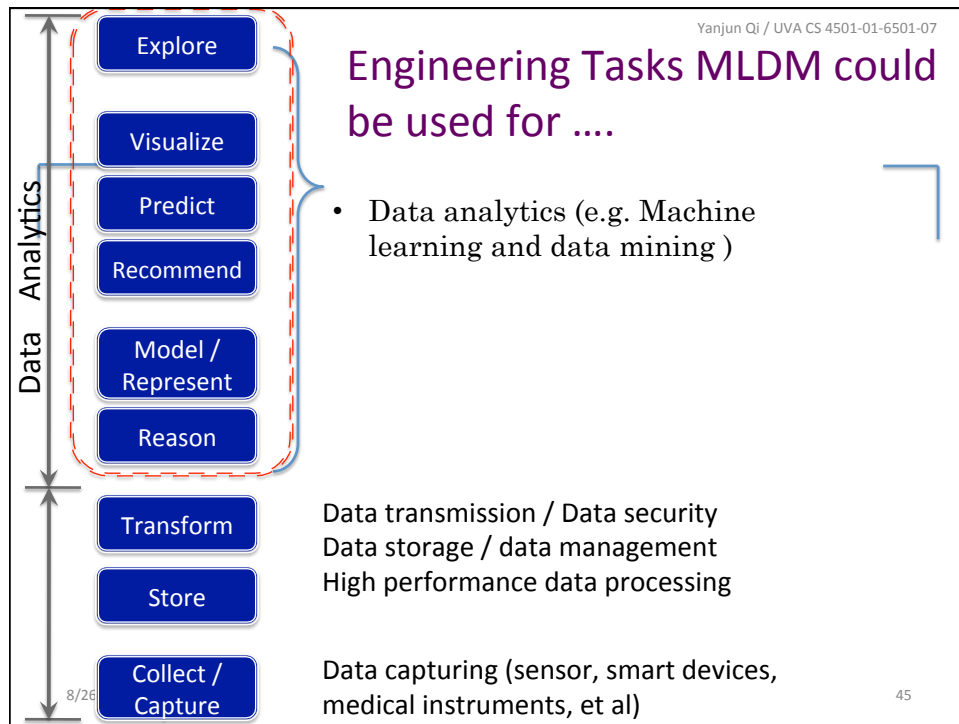
Prof. Mooney's slides⁴³

When to use Machine Learning (Adapt to / learn from data) ?

- 1. Extract knowledge from data
 - Relationships and correlations can be hidden within large amounts of data
 - The amount of knowledge available about certain tasks is simply too large for explicit encoding (e.g. rules) by humans
- 2. Learn tasks that are difficult to formalise
 - Hard to be defined well, except by examples
- 3. Create software that improves over time
 - New knowledge is constantly being discovered.
 - Rule or human encoding-based system is difficult to continuously re-design "by hand".

8/26/14

44



References

- Prof. Andrew Moore's slides
- Prof. Raymond J. Mooney's slides
- Prof. Alexander Gray's slides